GEOUJA System. University of Jaén at GEOCLEF 2007

José M. Perea-Ortega, Miguel A. García-Cumbreras, Manuel García-Vega, Arturo Montejo-Ráez
SINAI Group. Department of Computer Science. University of Jaén
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{jmperea,magc,mgarcia,amontejo}@ujaen.es

Abstract

This paper describes the second participation of the SINAI group of the University of Jaén in GeoCLEF 2007. We have developed a system different from the one presented in GeoCLEF 2006. Our architecture is made up of five main modules. The first one is the Information Retrieval Subsystem, that works with collections and queries in English and returns relevant documents for a query. The queries that are not in English are translated by the Translation Subsystem. All the queries are filtered by the Geo-Relation Finder Subsystem, that finds any spatial relation in the topic, and NER (Named Entities Recognition) Subsystem, that looks for any location in the topic. The most important module is the Geo-Relation Validator Subsystem, it applies some heuristics to filter documents recovered by the IR Subsystem. We have made several runs, combining these modules to resolve the monolingual and the bilingual tasks. The results obtained show that the heuristics applied are quite restrictive and therefore it must be generated new heuristics and to improve the definition of new rules to filter recovered documents.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Algorithms, Languages, Performance, Experimentation

Keywords

Information Retrieval, Geographic Information Retrieval, Named Entity Recognition, GeoCLEF

1 Introduction

The objective of GeoCLEF is to evaluate Geographical Information Retrieval (GIR) systems in tasks that involve both spatial and multilingual aspects. Given a multilingual statement describing a spatial user need (topic), the challenge is to find relevant documents from target collections in English, but with topics in English, Spanish, German or Portuguese [3]. This is our second participation in GeoCLEF, after the previous year[2].

In the last edition we studied the behavior of query expansion. The results obtained showed us that filtering improves precision and recall. For this reason, our system consists of five subsystems: Translation, Geographical Relations Finder, NER, Validator and Information Retrieval.

The most important one is the Validator module. The list of relevant documents is filtered using it: if a document doesn't pass the validation test it is removed from the list. Next section describes the whole system. Then, in the section 3, each module of the system is explained. Later on, results are described and finally, the conclusions about our participation in GeoCLEF 2007 are expounded.

2 System overview

We propose a Geographical Information Retrieval System that is made up of five related subsystems. These modules are explained in detail in the next section.

In our architecture we only worked with the English collection[1] and we have applied a off-line preprocess to it. This preprocess consists in using the English *stop-words* list, a named entity recognizer (NER) and the Porter *stemmer*[4]. The preprocessed data set will be indexed later using the IR Subsystem.

The translated query or topic proposed is indexed using the IR Subsystem too. If the language of topic is different from English, then it is translated by means of the Translation Subsystem. For each translated query to evaluate we labeled it with NER and *geo-relation* information. The Geo-Relation Finder Subsystem (GR Finder Subsystem) extracts spatial relations from the geographic query and the NER Subsystem recognizes named entities.

From the original English query the documents are recovered by the IR Subsystem that previously indexed the data collection. The NER information and the *geo-relation* components in relevant documents, and the NER locations from the geographic query are the input for Geo-Relation Validator Subsystem (GR Validator Subsystem), the most important module in our architecture.

In the GR Validator Subsystem we eliminate those relevant documents previously retrieved that do not agree with several rules. These rules are related to all the information that handles this module (locations and spatial relations from documents and geographic queries) and are explained in section 3.4. Figure 1 shows the proposed system architecture.

3 Subsystems description

3.1 Translation Subsystem

As translation module, we have used SINTRAM (SINai TRAnslation Module)[1]. This subsystem translates the queries from several languages into English. SINTRAM uses some on-line Machine Translators for each language pair and implements some heuristics to combine the different translations. After a complete research the best translators were found to be

- Systran for French, Italian and Portuguese. It is available at *http://www.systransoft.com*
- Prompt for Spanish. It is available at *http://translation2.paralink.com*

3.2 NER Subsystem

The main goal of NER Subsystem is to detect and recognize the entities appearing in the queries. We are only interested in geographical information, so we have just used *locations* detected by this NER Module. We have used the NER module of the GATE[2] toolkit. The location terms includes everything that is town, city, capital, country and even continent. The NER module adds *entity labels* to the topics with the found locations. An *entity label* example recognized in the title of the topic follows:

$$< en\_title\ position = "15"\ type = "LOC" > USA < /en\_title >$$

[1] English Los Angeles Times 94 (LA94) and English Glasgow Herald 95 (GH95)
[2] http://gate.ac.uk/

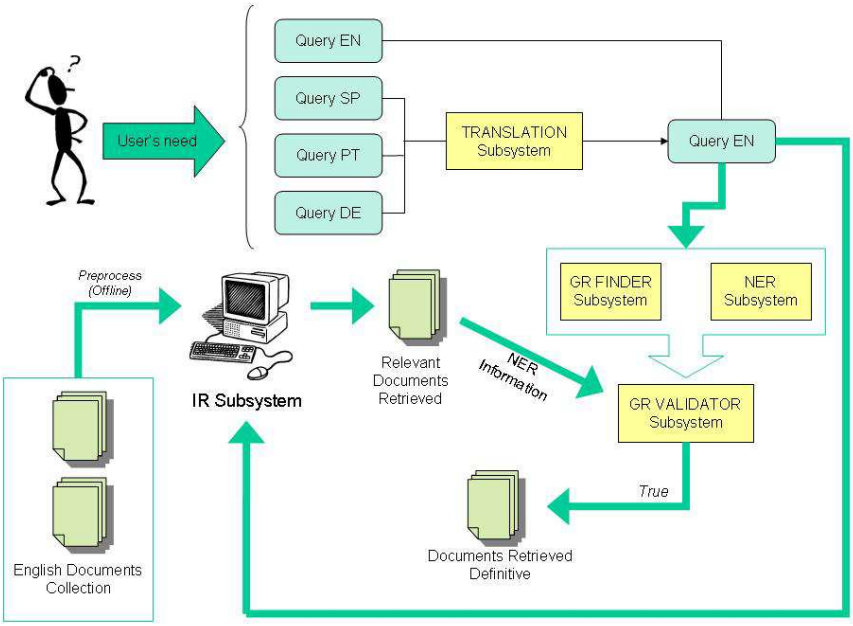

Figure 1: GEOUJA System architecture

where *position* is the position of the entity in the phrase. This value is greater than or equal to zero and we used it to know at any moment what locations and *geo-relations* are related to each other by proximity.

The basic operation of NER Subsystem is the following:

- The first step is the preprocessing phase. Each query is preprocessed using a tokenizer, a sentence splitter and a POS tagger. The NER Subsystem we have used needs this information in order to improves the named entity detection and recognition.
- The second is the detection of the geographical places. For this proposal we have use a Gazetteer, included also in GATE.

The NER Subsystem generates some topic labels, based on the original ones, adding the locations. These topic labels will be used later by the GR Validator Subsystem.

3.3 GR Finder Subsystem

The Geo-Relation Finder Subsystem is used to find the spatial relations in the geographic queries. This module makes use of four text files to store the *geo-relations* identified. Four *geo-relations* files exist because our system detects spatial relations of four words at the most. Some *geo-relations* examples are: *in, near, north of, next to, in or around, in the west of...*

The GR Finder module adds *geo-relation labels* to the topics with the found spatial relations. A *geo-relation label* example recognized in the title of the topic would be the following one:

$$< gr\_title\ position = "43" > near < /gr\_title >$$

where *position* is the position of the spatial relation in the phrase.

In this module we controlled a special *geo-relation* named *between*. For this case, the GR Finder Subsystem adds the two entities that this preposition relates. An example of the label that

this module adds for *description* label "*To be relevant documents describing oil or gas production between the UK and the European continent will be relevant*" is:

$$< gr\_desc\ position = "9" > between\ the;\ UK;\ European < /gr\_desc >$$

where we can see how both entities (*UK* and *European*) are added after the preposition, separated by a semicolon.

The basic operation of GR Finder Subsystem is the following:

- For each topic label (*title, desc* or *narr*) the subsystem looks for some spatial relation. It makes use of the text files that store the *geo-relations* that can detect.
- For each found spatial relation we verified that the word that comes next is an entity. For that reason it is necessary that the NER Subsystem is executed before.

Like NER Subsystem, the GR Finder Subsystem also generates topic labels, based on the original topic, adding the spatial relations. These topic labels (*entities* and *geo-relations*) will be used later by the GR Validator Subsystem. In the Figure 2 we can see an example of the text that generates this subsystem.

```
<?xml version="1.0" encoding="UTF-8"?>
<topics>
<top lang="en">
<num>10.2452/51-GC</num>
<title>Oil and gas extraction found between the UK and the Continent</title>
<desc>To be relevant documents describing oil or gas production between the
UK and the European continent will be relevant</desc>
<narr>Oil and gas fields in the North Sea will be relevant.</narr>
<en_title position="7" type="LOC">UK</en_title>
<en_desc position="11" type="LOC">UK</en_desc>
<en_desc position="14" type="MISC">European</en_desc>
<en_narr position="6" type="LOC">North Sea</en_narr>
<gr_title position="5">between the;UK;-</gr_title>
<gr_desc position="9">between the;UK;European</gr_desc>
<gr_narr position="4">in the</gr_narr>
</top>
<top lang="en">
<num>10.2452/52-GC</num>
<title>Crime near St Andrews</title>
<desc>To be relevant, documents must be about crimes occurring close to or
in St. Andrews.</desc>
<narr>Any event that refers to criminal dealings of some sort is relevant,
from thefts to corruption.</narr>
<en_title position="2" type="LOC">St Andrews</en_title>
<en_desc position="15" type="LOC">St Andrews</en_desc>
<gr_title position="1">near</gr_title>
<gr_desc position="9">in</gr_desc>
</top>
<top lang="en">
<num>10.2452/53-GC</num>
<title>Scientific research at east coast Scottish Universities</title>
<desc>For documents to be relevant, they must describe scientific research
conducted by a Scottish University located on the east coast of Scotland</desc>
<narr>Universities in Aberdeen, Dundee, St Andrews and Edinburgh wil be
considered relevant locations.</narr>
<en_desc position="21" type="LOC">Scotland</en_desc>
<en_narr position="2" type="LOC">Aberdeen</en_narr>
<en_narr position="13" type="MISC">Dundee St</en_narr>
<en_narr position="7" type="LOC">Edinburgh</en_narr>
<gr_desc position="20">of</gr_desc>
<gr_narr position="1">in</gr_narr>
</top>
```

Figure 2: Text example generated by GR Finder Subsystem

3.4 GR Validator Subsystem

This it is the most important module of our system. Its main goal is to discriminate what documents among the recovered ones by the IR Subsystem are valid.

In order to apply different heuristics, this module makes use of geographical data. This geo-information has been obtained from Geonames Gazetteer [3]. This module solves evaluations like:

[3]http://www.geonames.org/. Geonames geographical database contains over eight million geographical names and consists of 6.3 million unique features whereof 2.2 million populated places and 1.8 million alternate names

- Find the country name of a city.
- Find the latitude and longitude for a given location.
- Check if a city belongs to a certain country.
- Check if a location is to the north of another one.
- Calculate the distance from a location to another one.

Many heuristics can be applied with the former check points to make the validation of a document recovered by the IR Subsystem. The GR Validator Subsystem receives external information from IR Subsystem (entities from each document recovered) and from GR Finder and NER Subsystems (entities and spatial relations from each topic). This year we have used the following heuristics in our experiments:

1. For every entity appearing in query without an associated *geo-relation*, the system checks if this entity is present in documents recovered by IR Subsystem. The module discards a document whenever the number of entities found in topic with no associated *geo-relation* and not appearing in that document exceeds the fifty percent of the total of topic entities.
2. If the entity appearing in the topic has associated some spatial relation, the module checks if location is a continent, a country or a city. Depending on this location type for the query, the heuristics which we have followed in our experiments are the following ones:
   (a) If the location from a query is a continent or a country and its associated *geo-relation* is *in*, *on*, *at*, *from*, *of* or *along*, then the module checks if most of the entities of the document belong to that continent or country (at least fifty percent).
   (b) If the location from a query is a city and its associated spatial relation is *near*, *north of*, *south of*, *east of* or *west of*, the subsystem obtains the latitude and longitude information from Geonames Gazetteer about all locations from the document to be validated. The module will check if the geographic situation of each location is valid or not depending on the topic space relation.

For each heuristic to check, the system is adding or reducing points of a final score, depending on the result of that validation. A recovered document will be considered valid when the sum of all the scores when applying the heuristics to each entity of the document is greater than zero.

3.5 IR Subsystem

The information retrieval system that we have employed is Lemur[4]. It is a toolkit that supports indexing of large-scale text databases, the construction of simple language models for documents, queries, or sub-collections, and the implementation of retrieval systems based on language models as well as a variety of other retrieval models.

Previous to the index step, the English collection provided for GeoCLEF have been preprocessed, using the English *stop-words* list for meaningless terms removal, and the Porter *stemmer* [4] for suffix stripping. A NER has also been used to recognize possible entities in each document. Next, the English collection data set has been indexed using Lemur. After indexing the collection, each topic already translated is sent to Lemur. The relevant documents retrieved and their NER information will be used by the GR Validator Subsystem.

One parameter for each experiment is the weighting function, such as Okapi [5] or *TF.IDF*. Another is the use or not of Pseudo-Relevant Feedback (PRF) [6].

[4]http://www.lemurproject.org/. The toolkit is being developed as part of the Lemur Project, a collaboration between the Computer Science Department at the University of Massachusetts and the School of Computer Science at Carnegie Mellon University.

Experiment	Mean Average Precision	R-Precision
Sinai_ENEN_Exp1_fb_okapi	0.2605	0.2636
Sinai_ENEN_Exp1_fb_tfidf	0.1803	0.1858
Sinai_ENEN_Exp1_simple_okapi	0.2486	0.2624
Sinai_ENEN_Exp1_simple_tfidf	0.1777	0.1745
Sinai_ENEN_Exp2_fb_tfidf	0.1343	0.1656

Table 1: Summary of results for the monolingual task

4 Experiments and Results

SINAI[5] has participated in monolingual and bilingual tasks for GeoCLEF 2007 with a total of 26 experiments. In all experiments we have considered all tags from topics (*title*, *description* and *narrative*) as source for the information retrieval process.

Our baseline experiment consists in the Lemur retrieval on preprocessed collections (*stopper* and *stemmer*) without applying heuristics on relevant documents retrieved. This experiment has been applied in the monolingual and bilingual tasks.

The second experiment that we have made, consists in applying the heuristics that has been explained in section 3.4, on relevant documents retrieved by the IR Subsystem. This experiment also has been used in the monolingual and bilingual tasks.

4.1 Monolingual task

In the monolingual task we have participated with a total of 8 experiments: four about baseline experiment (*Exp1*) and other four about second experiment applying the heuristics introduced previously in this paper(*Exp2*). Some results are shown in Table 1.

Experiments named against an ending "*fb_okapi*" we have used Okapi with feedback as weighting function in the information retrieval process. Those ending with "*fb_tfidf*" indicate that we have applied *TF.IDF* with feedback. Also we have run experiments with Okapi but without feedback ("*simple_okapi*") and with *TF.IDF* but without feedback ("*simple_tfidf*").

4.2 Bilingual task

In the bilingual task we have participated with a total of 18 experiments: twelve about the baseline case (*Exp1*) and six applying our heuristics (*Exp2*). Some results are shown in Table 2.

For naming the experiments we have followed the same convention described in the previous section (see section 4.1). For German-English task we submitted six experiments in total. The string "*GEEN*" identifies them. For Portuguese-English task we submit six experiments too, identified by the string "*PTEN*". For Spanish-English task we submit six experiments in total (string "*SPEN*").

5 Conclusions and Future work

In this paper we have presented the experiments carried out in our second participation in the GeoCLEF campaign. The philosophy followed in this second experimental study has changed with respect to the approach presented last year. This year we have introduced a very restrictive system: we have tried to eliminate those documents recovered by the IR Subsystem that do not satisfy certain validation rules. However, the previous year we were centered in increasing the queries, expanding them with entities and thesauri information in order to improve retrieval effectiveness.

The results obtained the previous year showed that query expansion does not improve in general the quality of the information retrieval process. The results of this year shown that the documents

[5] http://sinai.ujaen.es

Experiment	Mean Average Precision	R-Precision
Sinai_GEEN_Exp1_fb_okapi	0.0686	0.0704
Sinai_PTEN_Exp1_fb_okapi	0.1568	0.1519
Sinai_SPEN_Exp1_fb_okapi	0.2362	0.2238
Sinai_GEEN_Exp1_fb_tfidf	0.0572	0.0606
Sinai_PTEN_Exp1_fb_tfidf	0.1080	0.1133
Sinai_SPEN_Exp1_fb_tfidf	0.1511	0.1533
Sinai_GEEN_Exp1_simple_okapi	0.0484	0.0569
Sinai_PTEN_Exp1_simple_okapi	0.1544	0.1525
Sinai_SPEN_Exp1_simple_okapi	0.2310	0.2476
Sinai_GEEN_Exp1_simple_tfidf	0.0435	0.0420
Sinai_PTEN_Exp1_simple_tfidf	0.1053	0.1117
Sinai_SPEN_Exp1_simple_tfidf	0.1447	0.1513
Sinai_PTEN_Exp2_fb_tfidf	0.0695	0.1074

Table 2: Summary of results for the bilingual task

that have been recovered are valid but the GR Validator Subsystem has filtered some ones that must not have eliminated.

For the future, we will try to add more heuristics to the GR Validator Subsystem making use of Geonames Gazetteer. Also we will define more precise rules so that the system is less restrictive for the selection of recovered documents. Finally, we will also explore a larger number of retrieved documents by the IR Subsystem, in the aim of providing a larger variety of documents to be checked by the GR Validator Subsystem.

6 Acknowledgments

This work has been supported by Spanish Government (MCYT) with grant TIN2006-15265-C06-03.

References

[1] Miguel A. García-Cumbreras, L. Alfonso Ureña-López, Fernando Martínez Santiago, and José M. Perea-Ortega. Bruja system. the university of jaén at the spanish task of qa@clef 2006. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2006)*, 2006.

[2] Manuel García-Vega, Miguel A. García-Cumbreras, L.A. Ureña-López, and José M. Perea-Ortega. Geouja system. the first participation of the university of jaén at geoclef 2006. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2006)*, 2006.

[3] Fredric Gey, Ray Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, and Paulo Rocha. Geoclef 2006: the clef 2006 cross-language geographic information retrieval track overview. In *Proceedings of the Cross Language Evaluation Forum (CLEF 2006)*, 2006.

[4] M.F. Porter. An algorithm for suffix stripping. In *Program 14*, pages 130–137, 1980.

[5] S.E. Robertson and S.Walker. Okapi-Keenbow at TREC-8. In *Proceedings of the 8th Text Retrieval Conference TREC-8, NIST Special Publication 500-246*, pages 151–162, 1999.

[6] G. Salton and G. Buckley. Improving retrieval performance by relevance feedback. *Journal of American Society for Information Sciences*, 21:288–297, 1990.